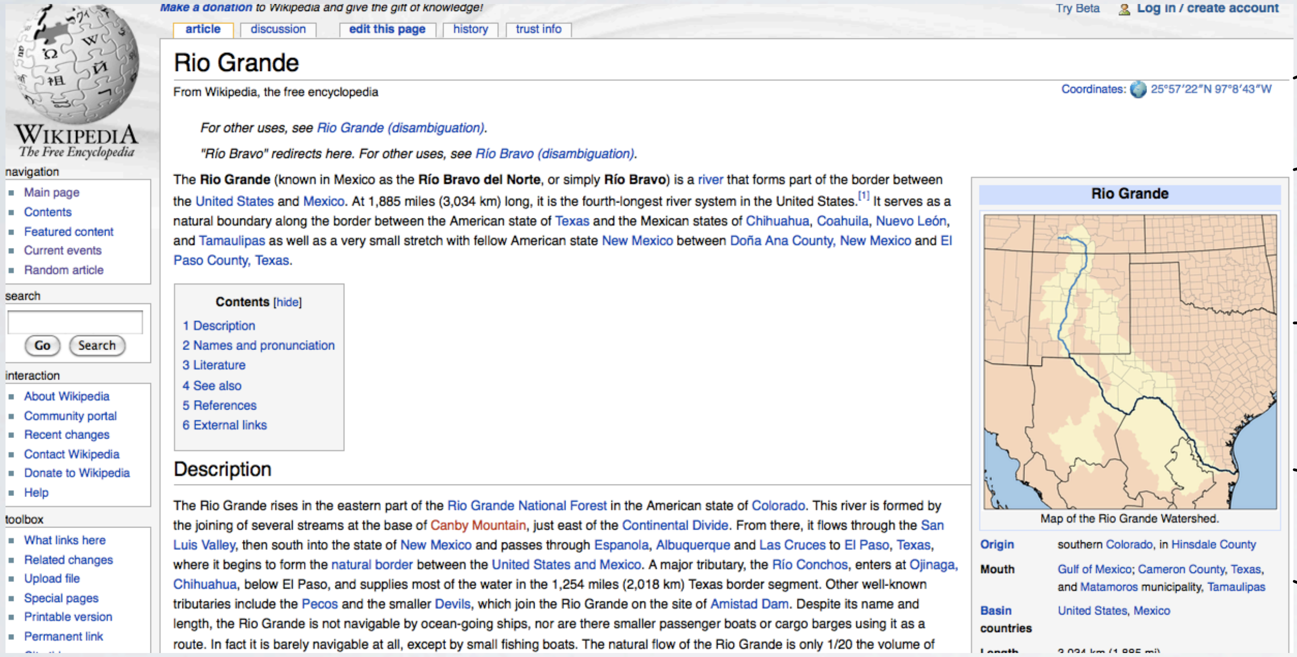
SCRAWL: A SEMANTIC CRAWLER FOR THE WIKIPEDIA AND BEYOND

Ian Pye, Luca de Alfaro
Shelly Spearing, Jorge Roman

THE PROBLEM



Make a donation to Wikipedia and give the gift of knowledge!

Try Beta Log in / create account

article | discussion | edit this page | history | trust info

WIKIPEDIA
The Free Encyclopedia

navigation
- Main page
- Contents
- Featured content
- Current events
- Random article

search

Go Search

interaction
- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

toolbox
- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link

Rio Grande

From Wikipedia, the free encyclopedia

Coordinates: 25°57′22″N 97°8′43″W

For other uses, see Rio Grande (disambiguation).

"Río Bravo" redirects here. For other uses, see Río Bravo (disambiguation).

The **Rio Grande** (known in Mexico as the **Río Bravo del Norte**, or simply **Río Bravo**) is a river that forms part of the border between the United States and Mexico. At 1,885 miles (3,034 km) long, it is the fourth-longest river system in the United States.[1] It serves as a natural boundary along the border between the American state of Texas and the Mexican states of Chihuahua, Coahuila, Nuevo León, and Tamaulipas as well as a very small stretch with fellow American state New Mexico between Doña Ana County, New Mexico and El Paso County, Texas.

Contents [hide]
1 Description
2 Names and pronunciation
3 Literature
4 See also
5 References
6 External links

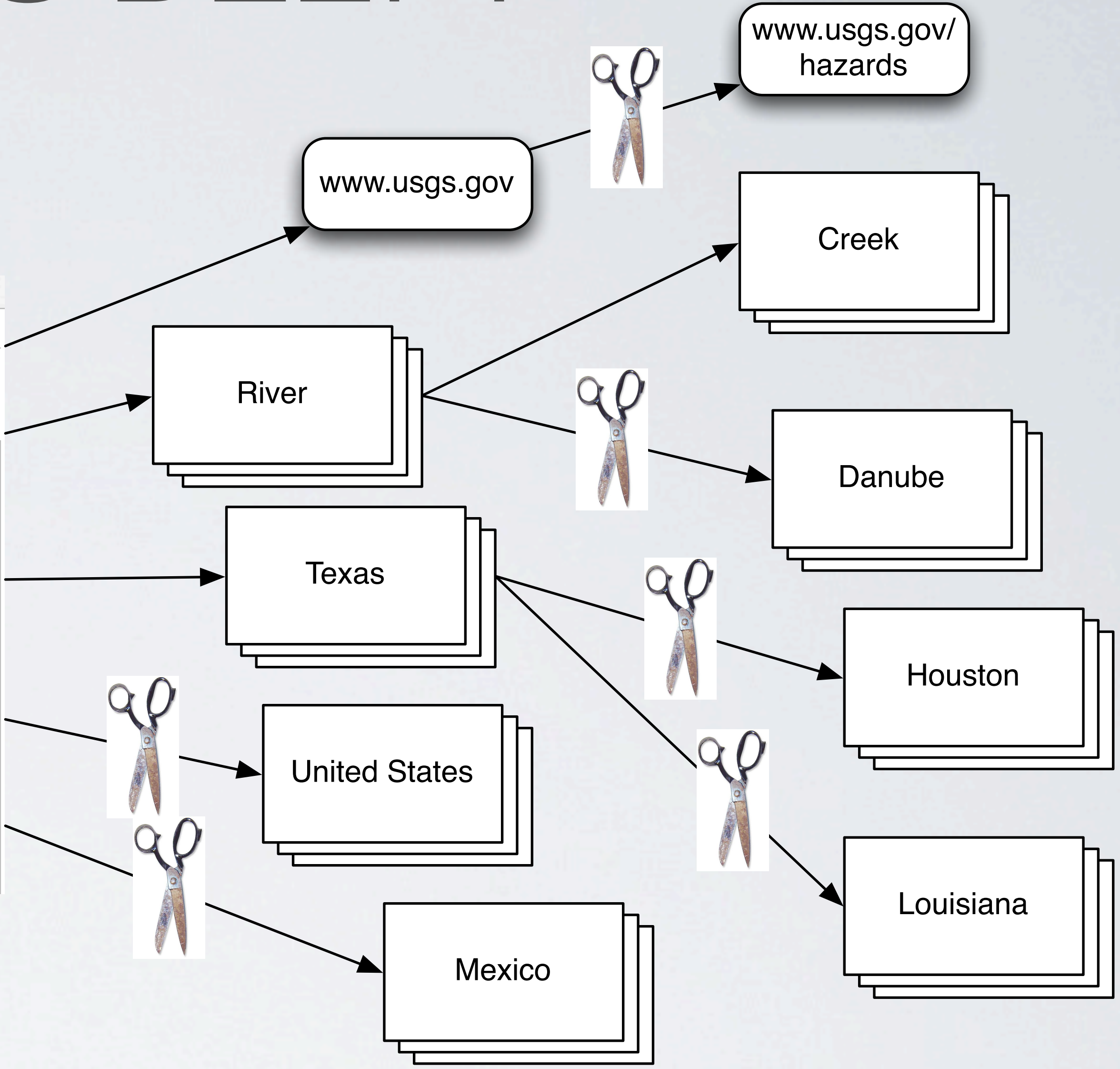
Description

The Rio Grande rises in the eastern part of the Rio Grande National Forest in the American state of Colorado. This river is formed by the joining of several streams at the base of Canby Mountain, just east of the Continental Divide. From there, it flows through the San Luis Valley, then south into the state of New Mexico and passes through Espanola, Albuquerque and Las Cruces to El Paso, Texas, where it begins to form the natural border between the United States and Mexico. A major tributary, the Río Conchos, enters at Ojinaga, Chihuahua, below El Paso, and supplies most of the water in the 1,254 miles (2,018 km) Texas border segment. Other well-known tributaries include the Pecos and the smaller Devils, which join the Rio Grande on the site of Amistad Dam. Despite its name and length, the Rio Grande is not navigable by ocean-going ships, nor are there smaller passenger boats or cargo barges using it as a route. In fact it is barely navigable at all, except by small fishing boats. The natural flow of the Rio Grande is only 1/20 the volume of

Rio Grande

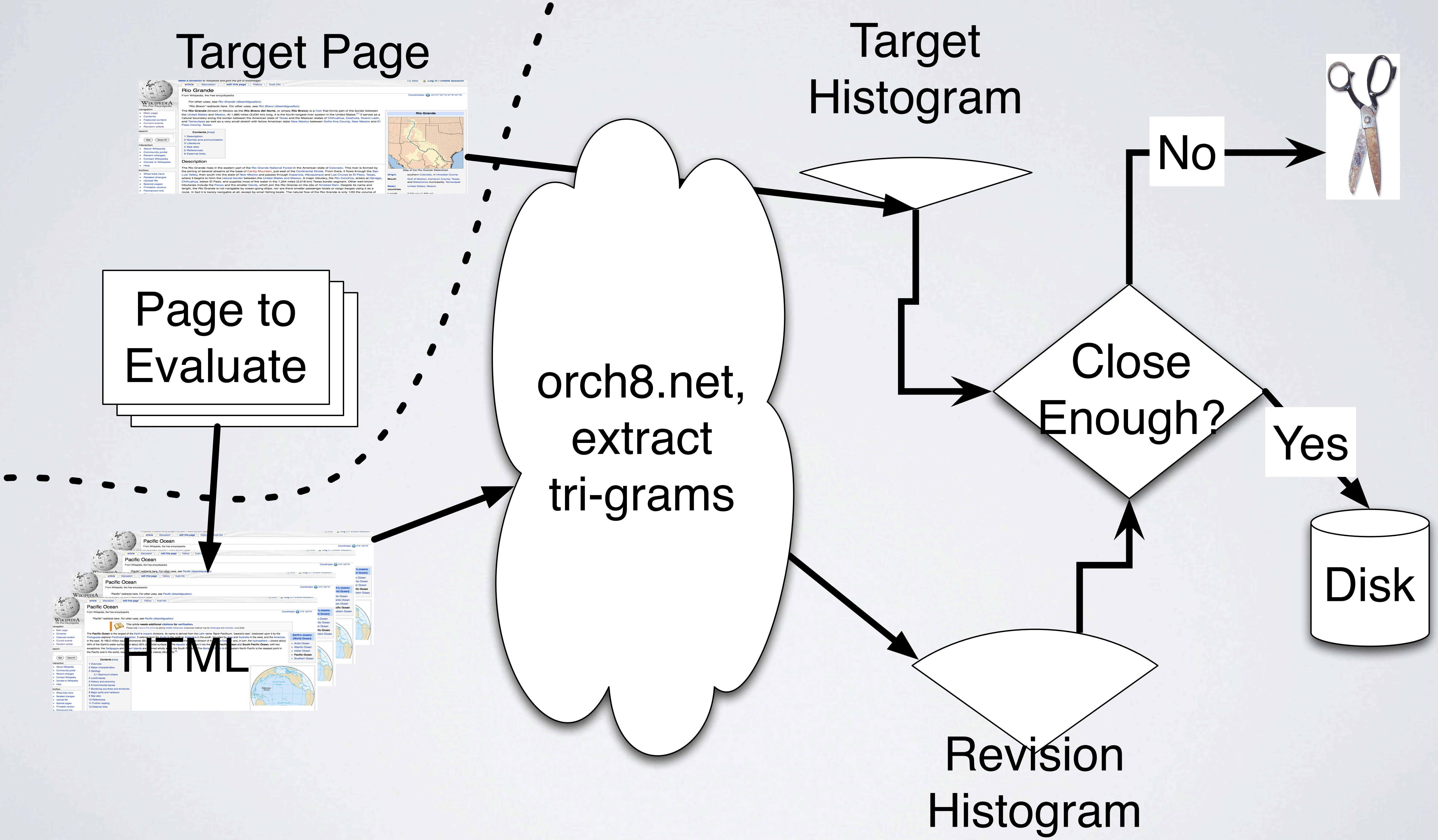
Map of the Rio Grande Watershed.

Origin	southern Colorado, in Hinsdale County
Mouth	Gulf of Mexico; Cameron County, Texas, and Matamoros municipality, Tamaulipas
Basin countries	United States, Mexico
Length	3,034 km (1,885 mi)



The Wikipedia is too big! We just want to look at the (potentially) interesting parts.

OUR SOLUTION



Target Page

Page to Evaluate

HTML

orch8.net, extract tri-grams

Target Histogram

Revision Histogram

Close Enough?

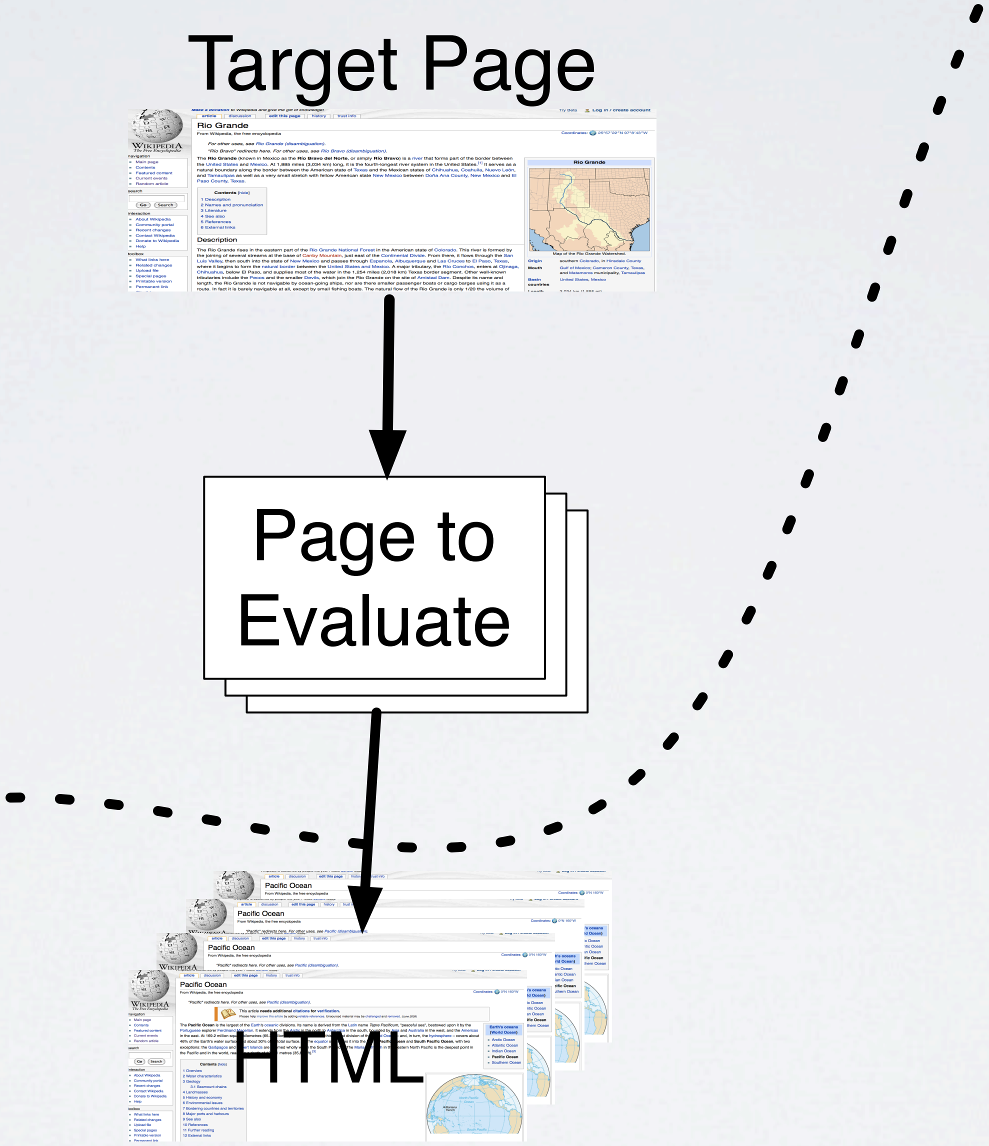
No

Yes

Disk

OUR SOLUTION

1) Download and Render to HTML



Target Page

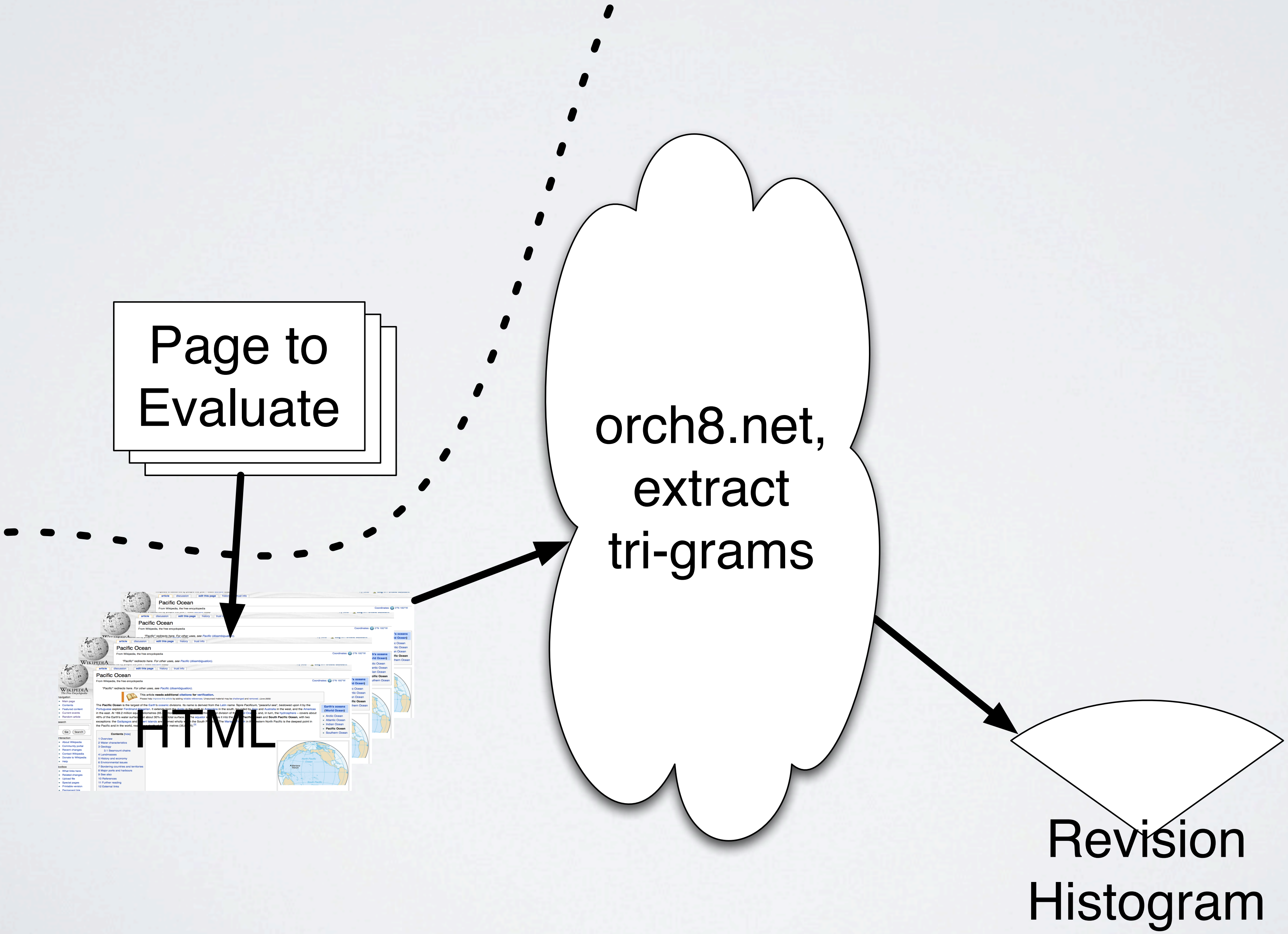
Page to Evaluate

HTML

OUR SOLUTION

1) Download and Render to HTML

2) From HTML to Semantic Digest



Page to Evaluate

HTML

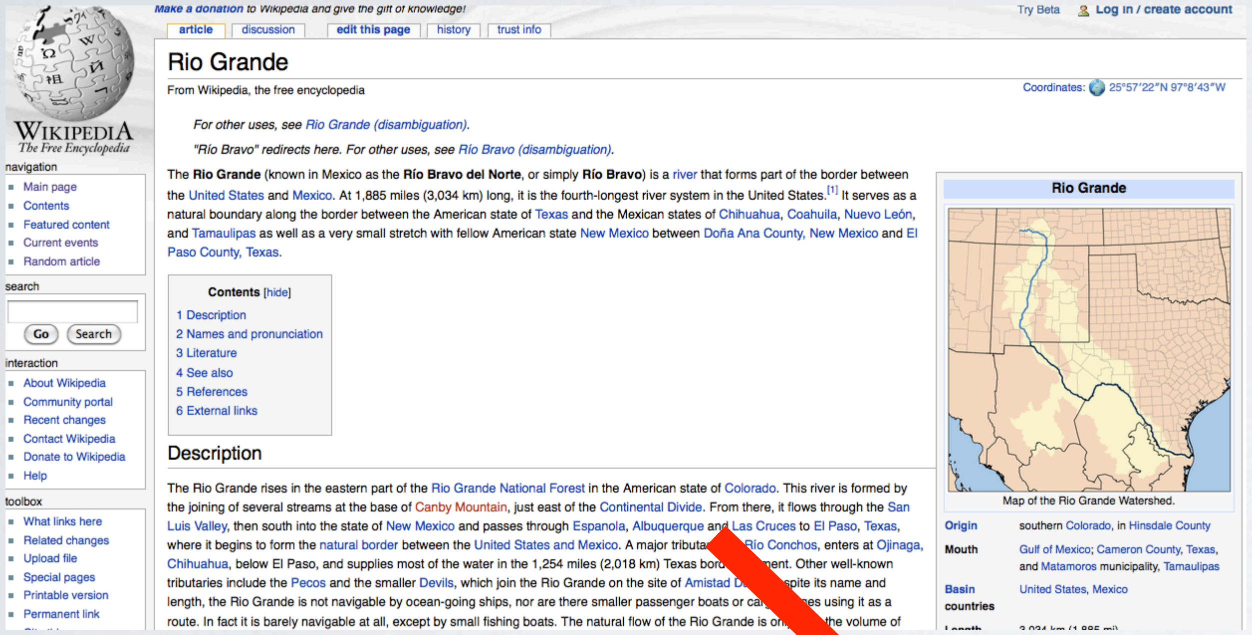
orch8.net, extract tri-grams

Revision Histogram

OUR SOLUTION

1) Download and Render to HTML

2) From HTML to Semantic Digest



Make a donation to Wikipedia and give the gift of knowledge!

Try Beta Log in / create account

article | discussion | edit this page | history | trust info

WIKIPEDIA
The Free Encyclopedia

navigation
- Main page
- Contents
- Featured content
- Current events
- Random article

search

Go Search

interaction
- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

toolbox
- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link

Rio Grande

From Wikipedia, the free encyclopedia

Coordinates: 25°57′22″N 97°8′43″W

For other uses, see Rio Grande (disambiguation).

"Río Bravo" redirects here. For other uses, see Río Bravo (disambiguation).

The **Rio Grande** (known in Mexico as the **Río Bravo del Norte**, or simply **Río Bravo**) is a river that forms part of the border between the United States and Mexico. At 1,885 miles (3,034 km) long, it is the fourth-longest river system in the United States.[1] It serves as a natural boundary along the border between the American state of Texas and the Mexican states of Chihuahua, Coahuila, Nuevo León, and Tamaulipas as well as a very small stretch with fellow American state New Mexico between Doña Ana County, New Mexico and El Paso County, Texas.

Contents [hide]
1 Description
2 Names and pronunciation
3 Literature
4 See also
5 References
6 External links

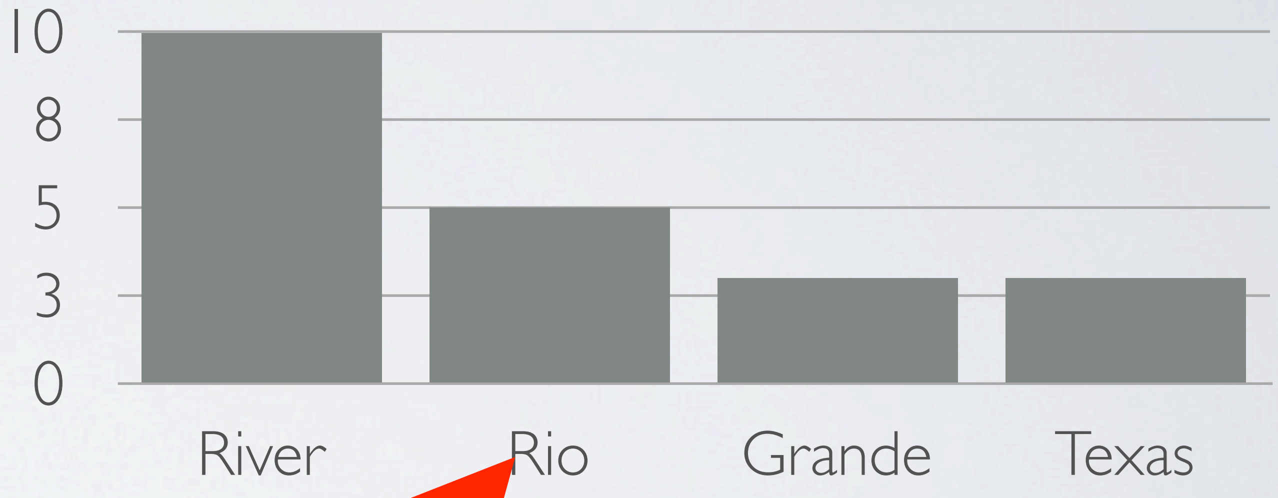
Description

The Rio Grande rises in the eastern part of the Rio Grande National Forest in the American state of Colorado. This river is formed by the joining of several streams at the base of Canby Mountain, just east of the Continental Divide. From there, it flows through the San Luis Valley, then south into the state of New Mexico and passes through Espanola, Albuquerque and Las Cruces to El Paso, Texas, where it begins to form the natural border between the United States and Mexico. A major tributa Río Conchos, enters at Ojinaga, Chihuahua, below El Paso, and supplies most of the water in the 1,254 miles (2,018 km) Texas bord ment. Other well-known tributaries include the Pecos and the smaller Devils, which join the Rio Grande on the site of Amistad D spite its name and length, the Rio Grande is not navigable by ocean-going ships, nor are there smaller passenger boats or ca es using it as a route. In fact it is barely navigable at all, except by small fishing boats. The natural flow of the Rio Grande is o the volume of

Rio Grande

Map of the Rio Grande Watershed.

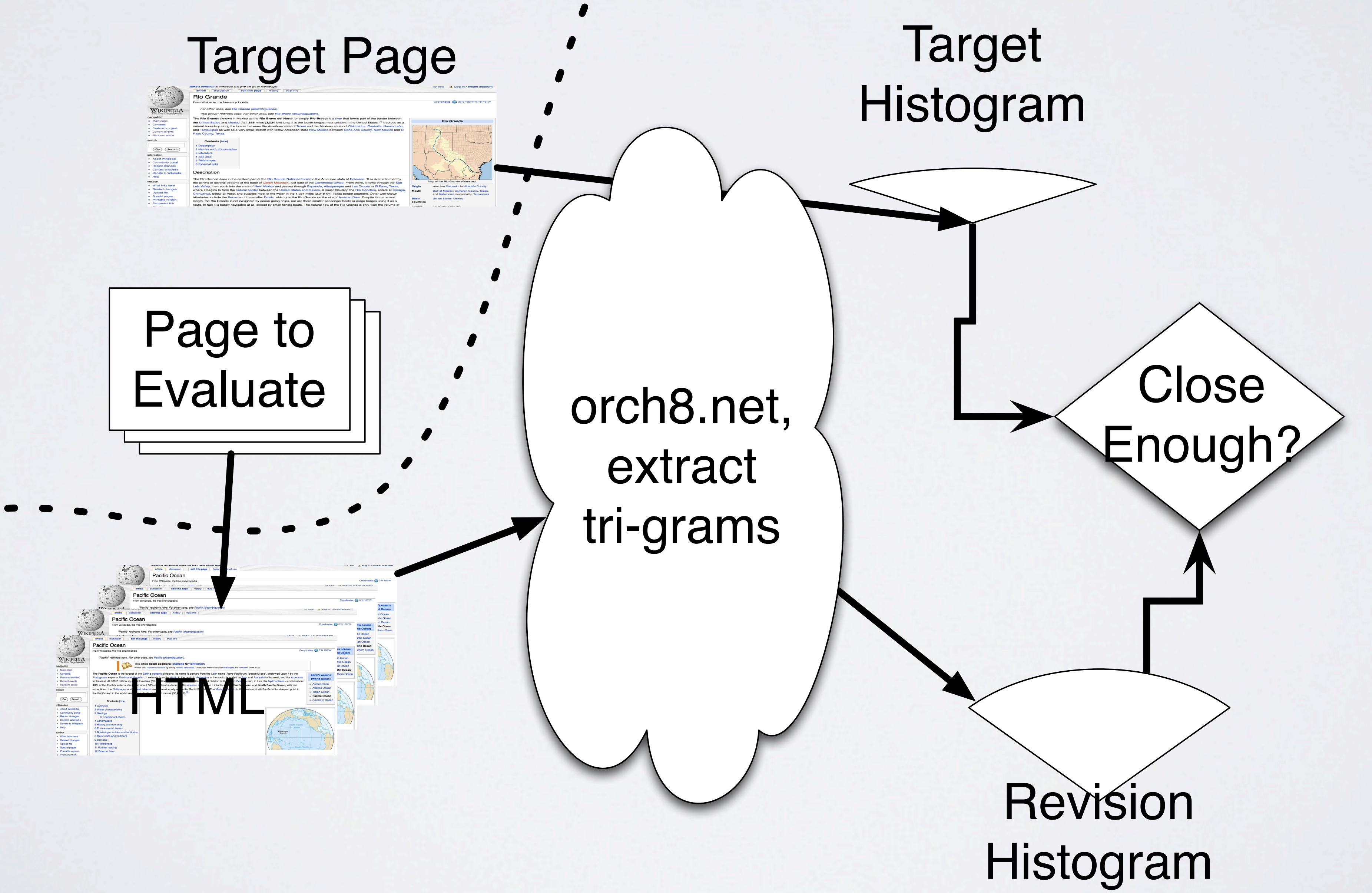
Origin	southern Colorado, in Hinsdale County
Mouth	Gulf of Mexico; Cameron County, Texas, and Matamoros municipality, Tamaulipas
Basin countries	United States, Mexico
Length	3,034 km (1,885 mi)



Rio Grande River
Texas State River
Big Bend Park

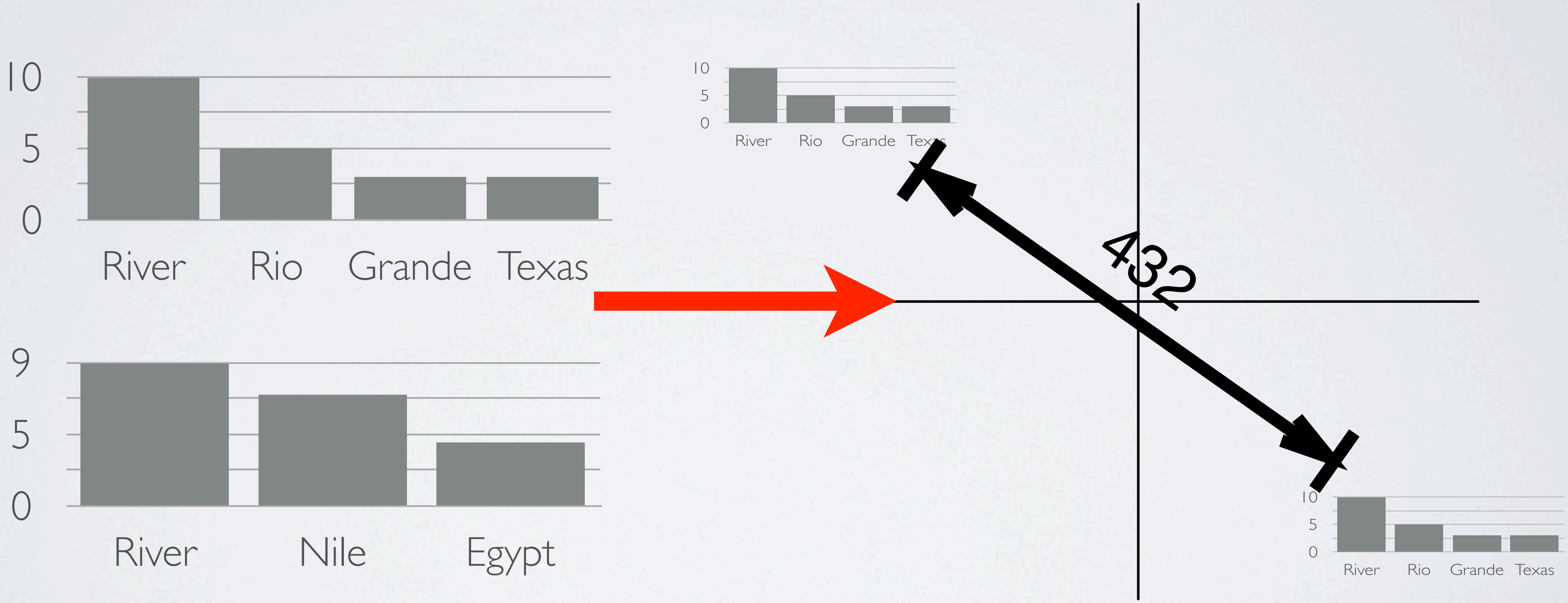
OUR SOLUTION

1) Download and Render to HTML
2) From HTML to Semantic Digest
3) Euclidian Distance from the Target



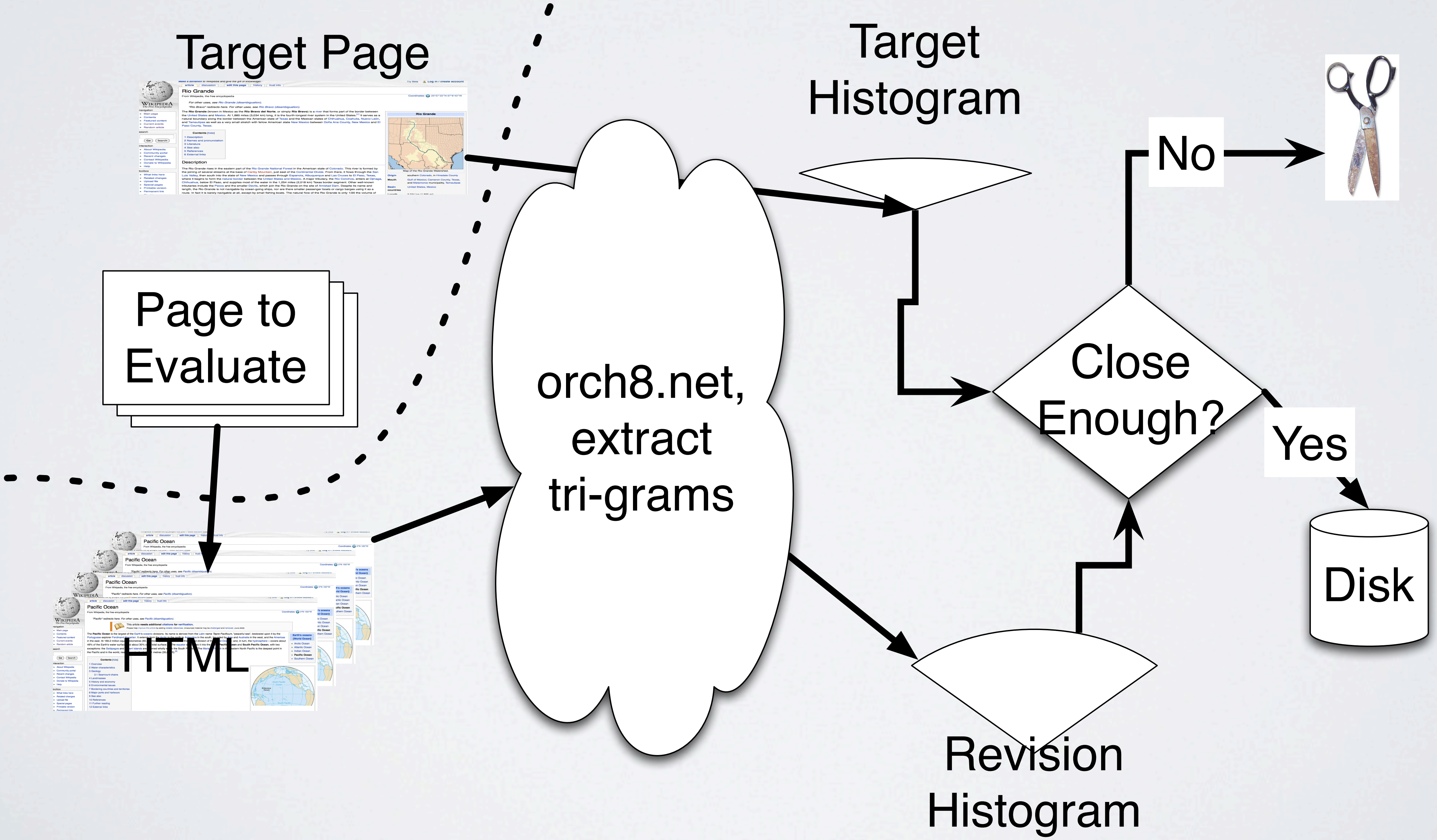
OUR SOLUTION

1) Download and Render to HTML
2) From HTML to Semantic Digest
3) Euclidian Distance from the Target



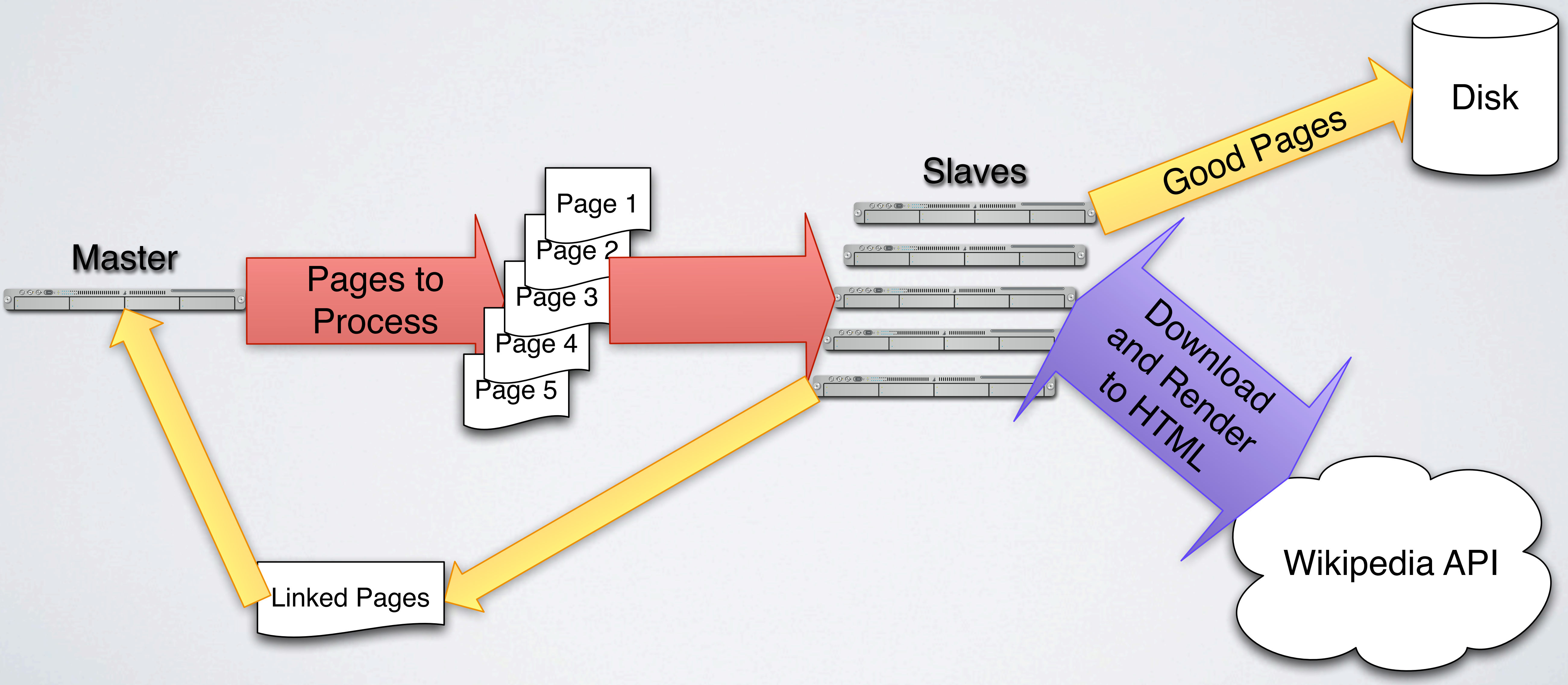
OUR SOLUTION

1) Download and Render to HTML
2) From HTML to Semantic Digest
3) Euclidian Distance from the Target
4) Extract Links and Recurse If Distance < Max



DO THIS IN PARALLEL

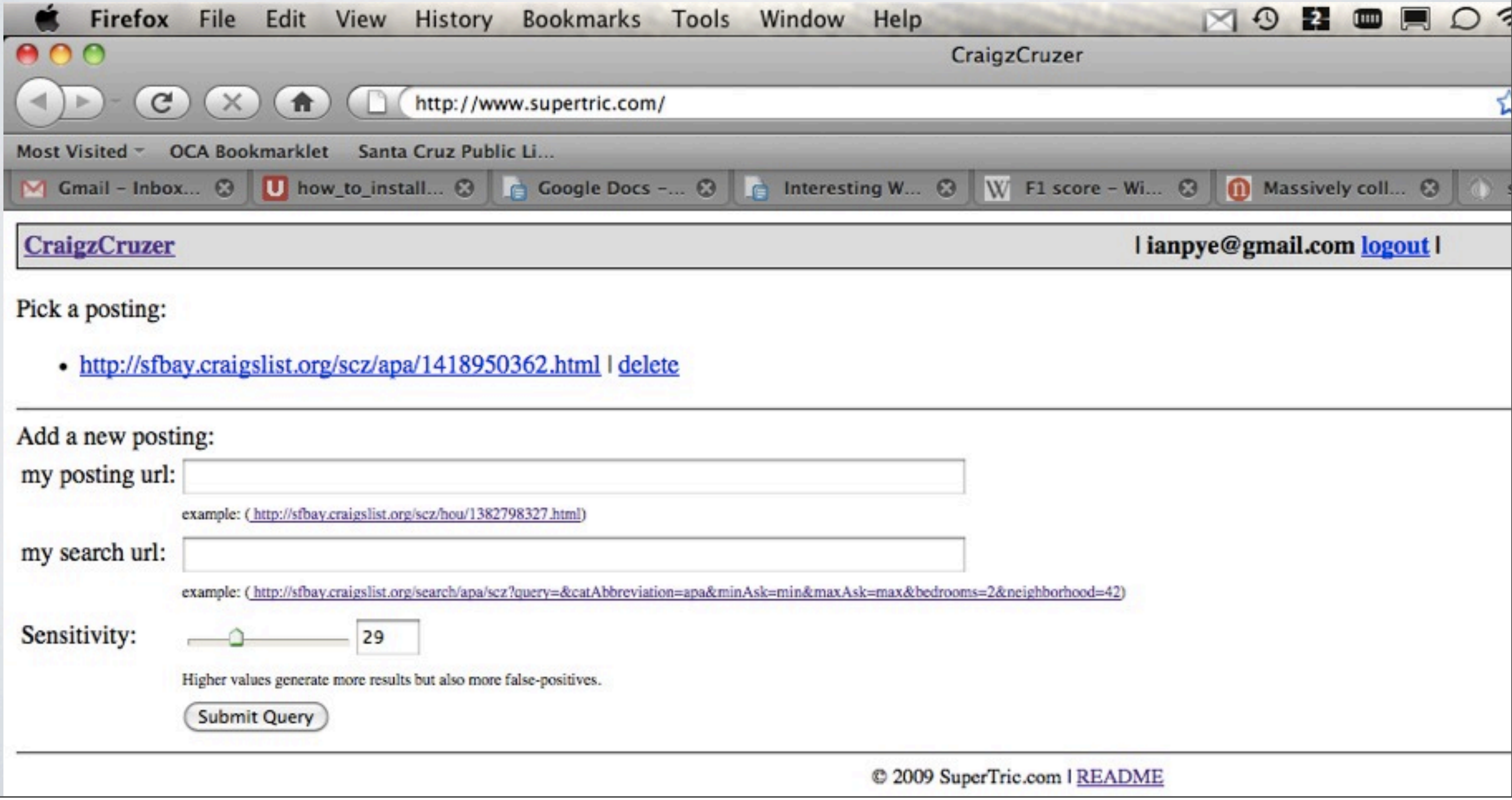
OpenMPI allows us to run in a page-wise parallel fashion on a cluster.



TESTING PLATFORM:

CRAIGZCRUZER

WWW.SUPERTRIC.COM



CraigzCruzer | ianpye@gmail.com logout |

Pick a posting:

- http://sfbay.craigslist.org/scz/apa/1418950362.html | delete

Add a new posting:

my posting url:

example: (http://sfbay.craigslist.org/scz/hou/1382798327.html)

my search url:

example: (http://sfbay.craigslist.org/search/apa/scz?query=&catAbbreviation=apa&minAsk=min&maxAsk=max&bedrooms=2&neighborhood=42)

Sensitivity: 29

Higher values generate more results but also more false-positives.

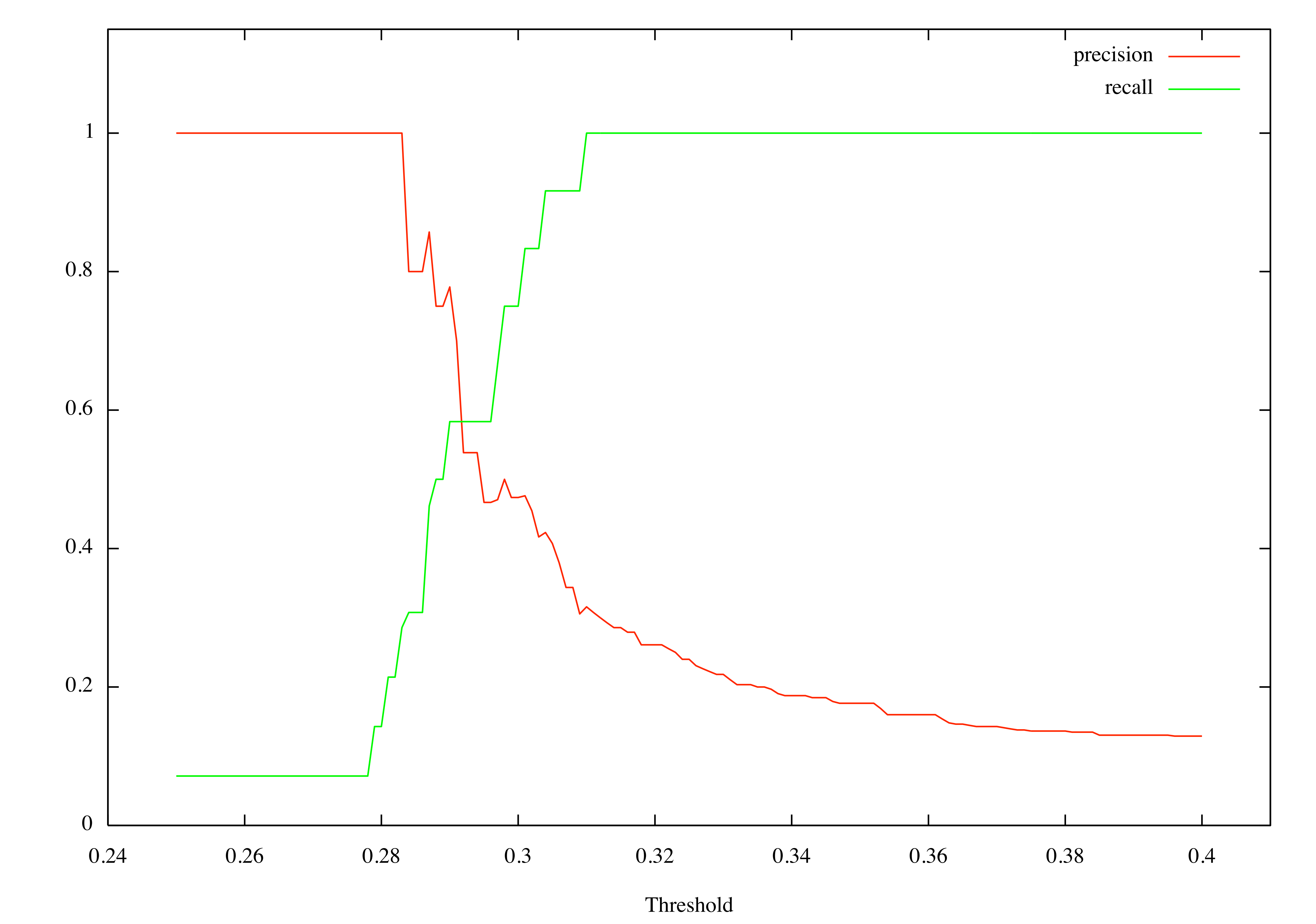
© 2009 SuperTric.com | README

HOW ACCURATE IS IT?

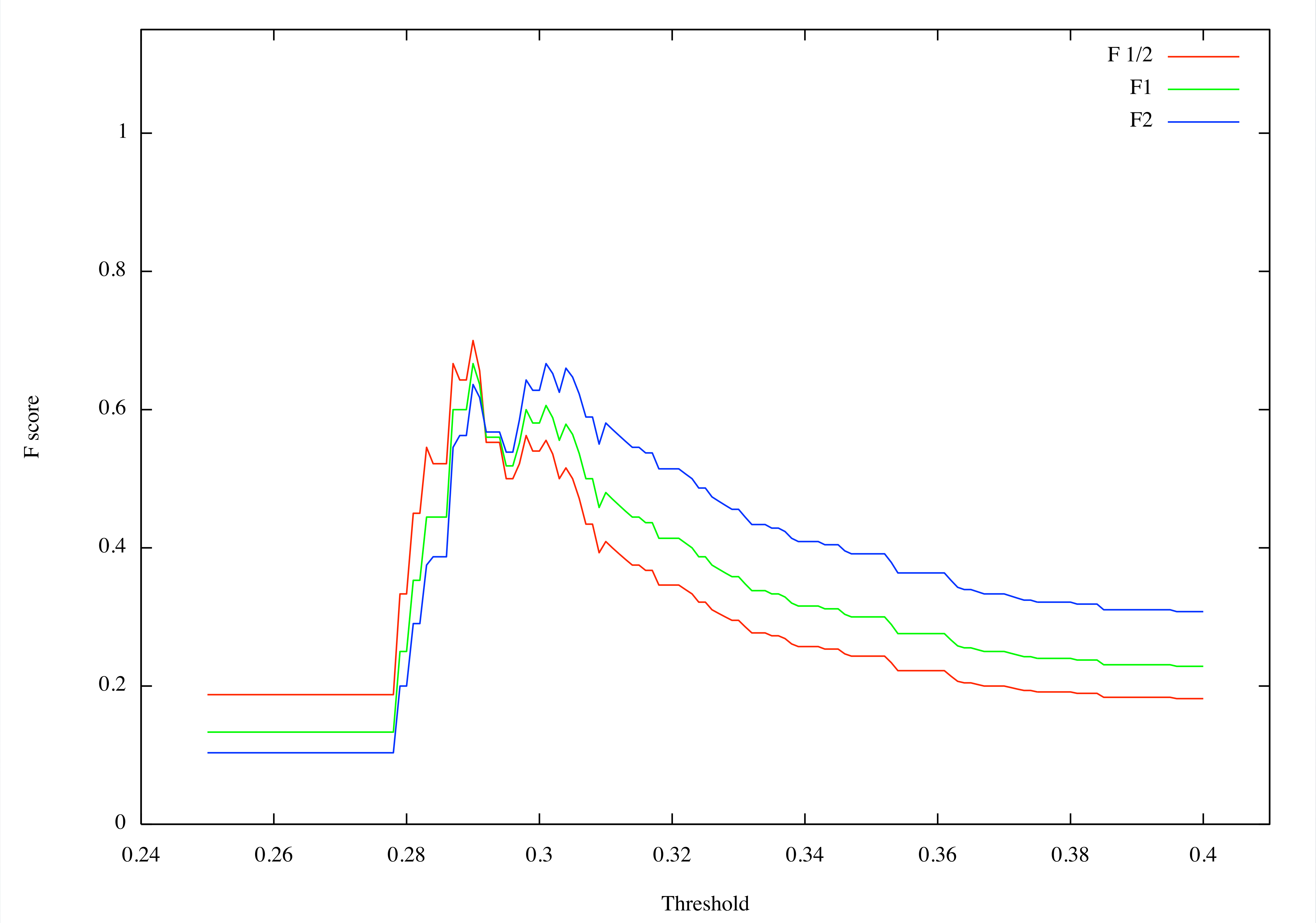
METRICS (OR, A SHORT DIGRESSION INTO IR)

- Precision: the ratio of true positives to false positives
- Recall: the ratio of found positives to all positives
- F_β: How much to weight recall vs precision?

RESULTS: PRECISION AND RECALL



RESULTS: F_β



CONCLUSIONS

- SCrawl is a tool to extract what text MIGHT be interesting.
- Works though semantic meaning.
- Achieve high recall without sacrificing too much precision.
- Parallel processing for large datasets.